Knowledge Containers tagged to Taxonomies

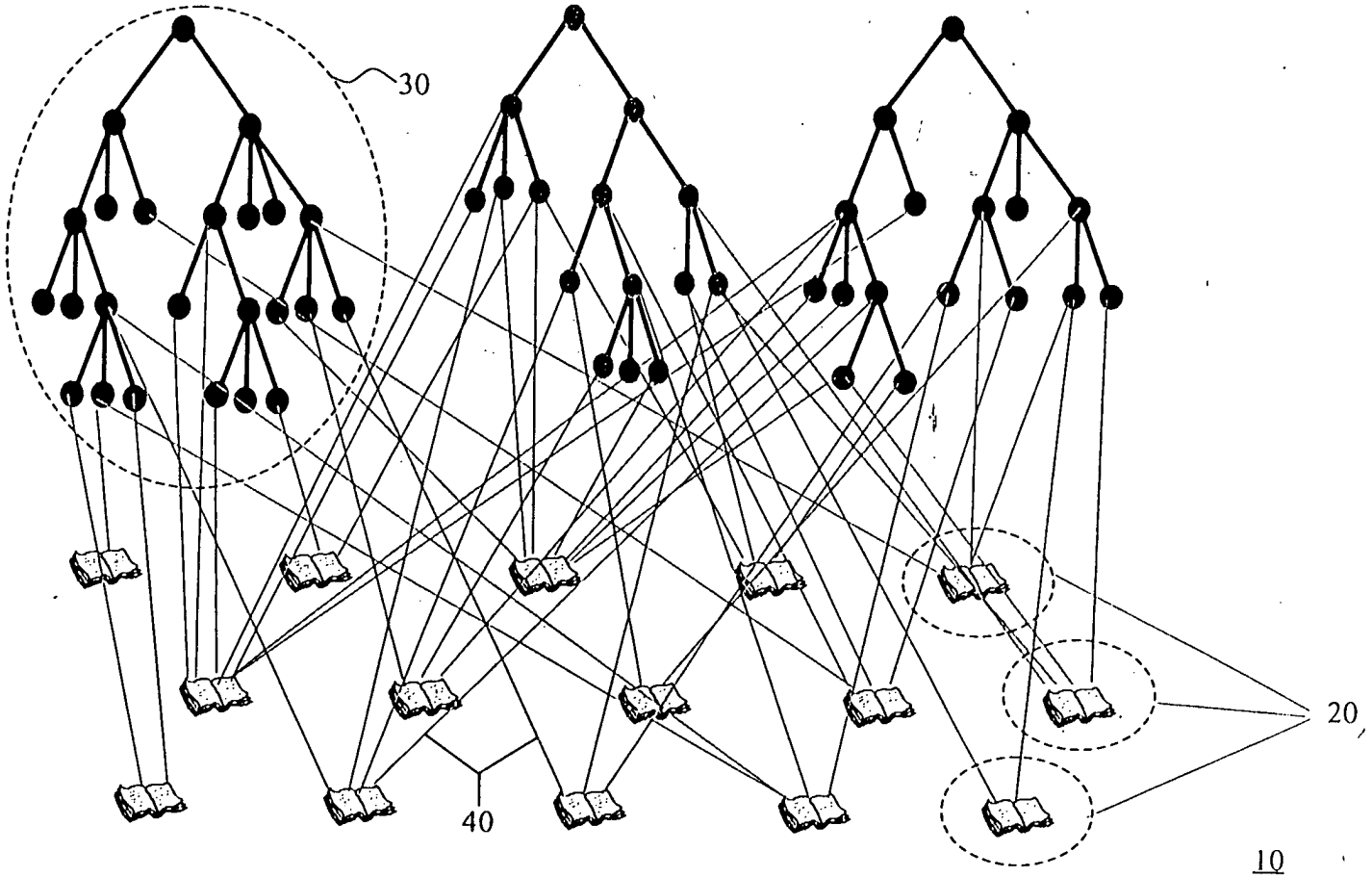
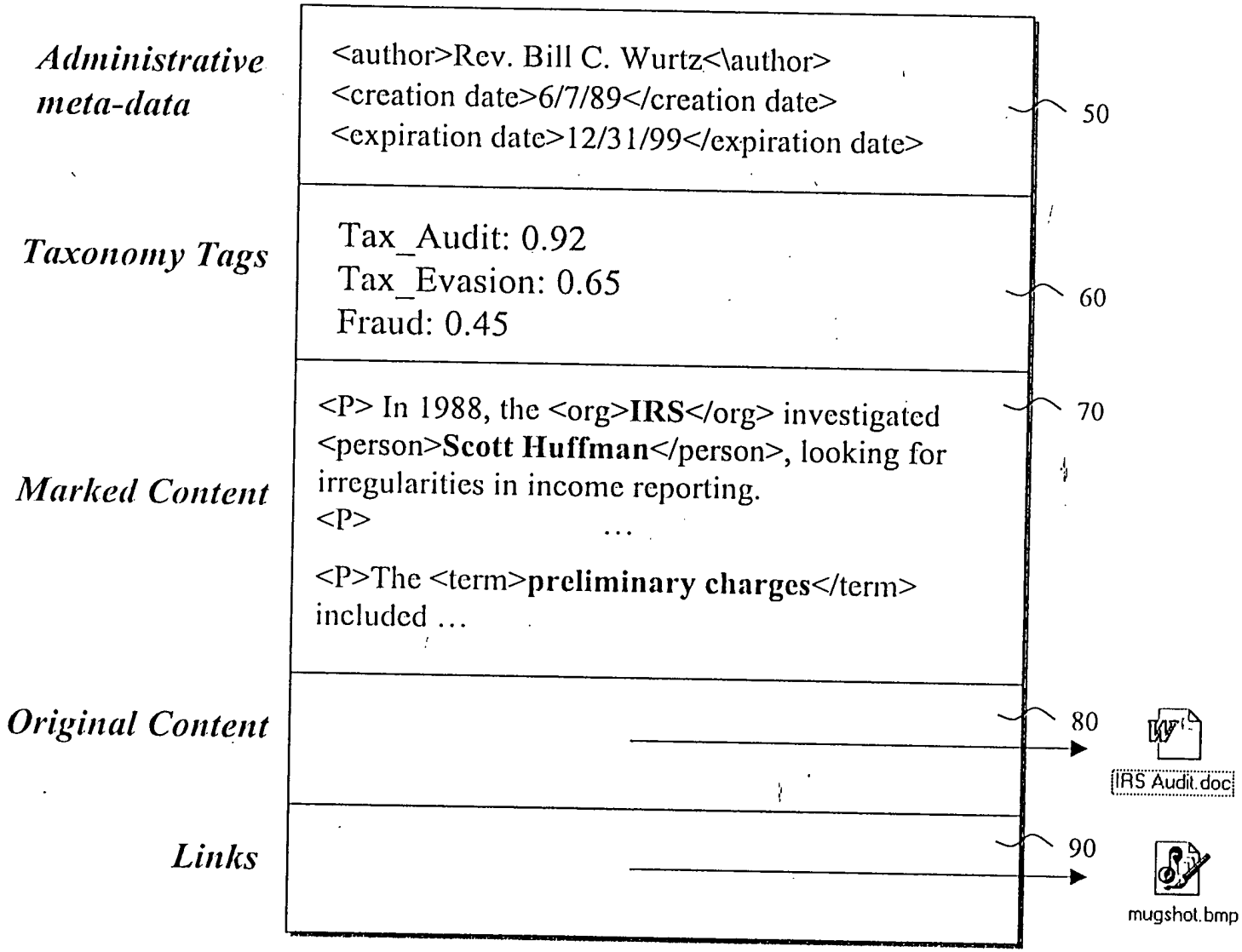


FIG. 1

A Knowledge Container



20

FIG. 2

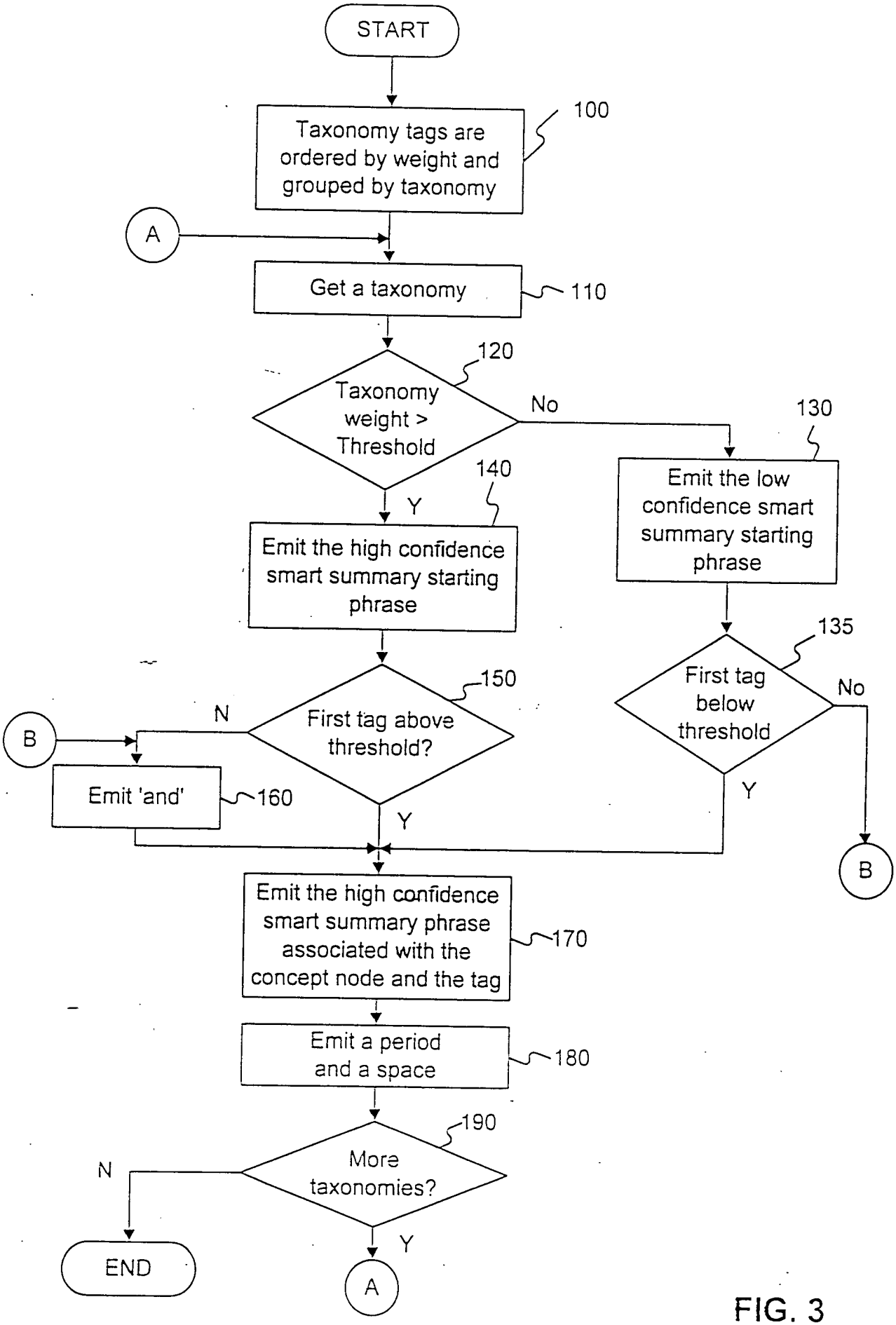
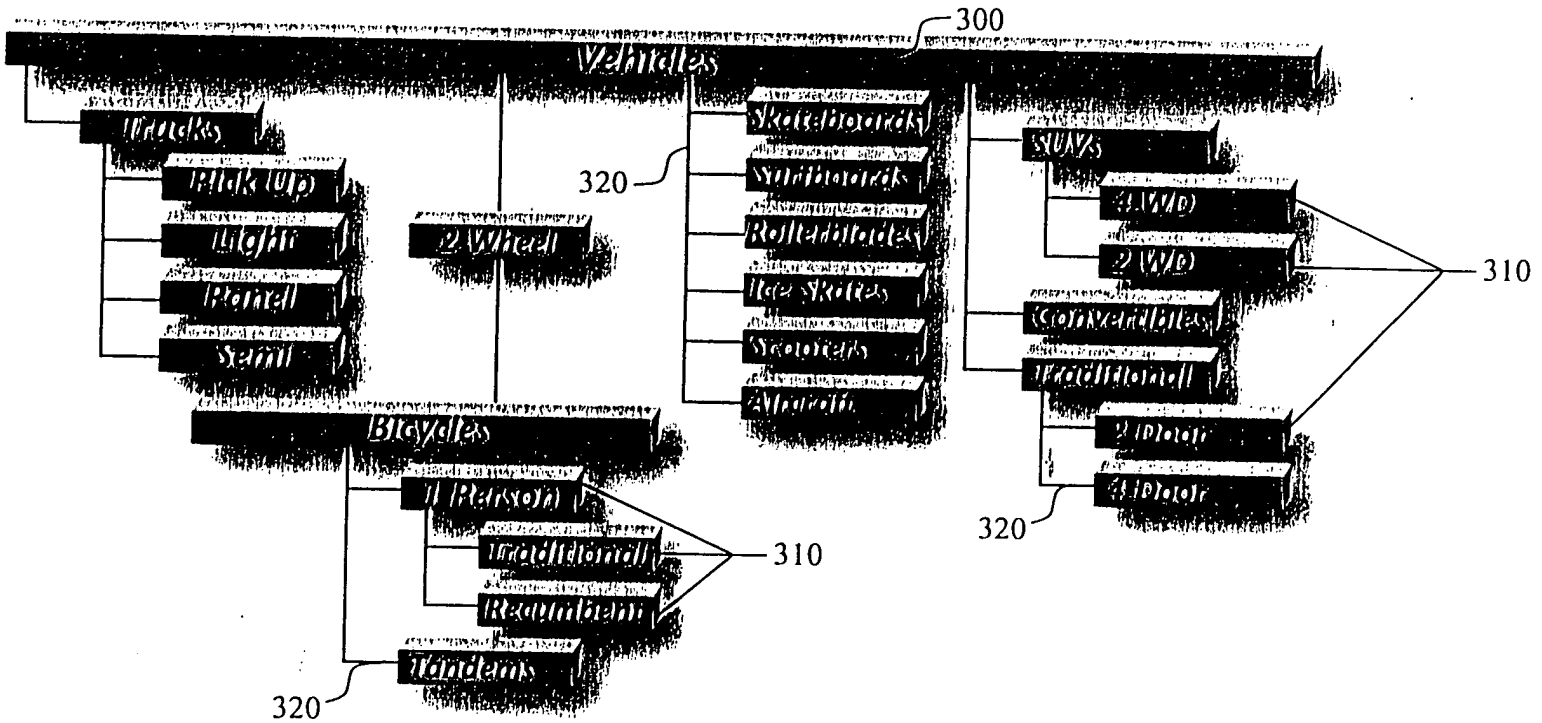

START
Taxonomy tags are ordered by weight and grouped by taxonomy
100
A
Get a taxonomy
110
120
Taxonomy weight > Threshold
No
130
Y
140
Emit the high confidence smart summary starting phrase
Emit the low confidence smart summary starting phrase
135
First tag below threshold
No
Y
150
First tag above threshold?
N
B
Emit 'and'
160
Y
B
Emit the high confidence smart summary phrase associated with the concept node and the tag
170
Emit a period and a space
180
190
More taxonomies?
N
Y
END
A

FIG. 3

Representative Taxonomy showing Types of Vehicles



30

FIG. 4

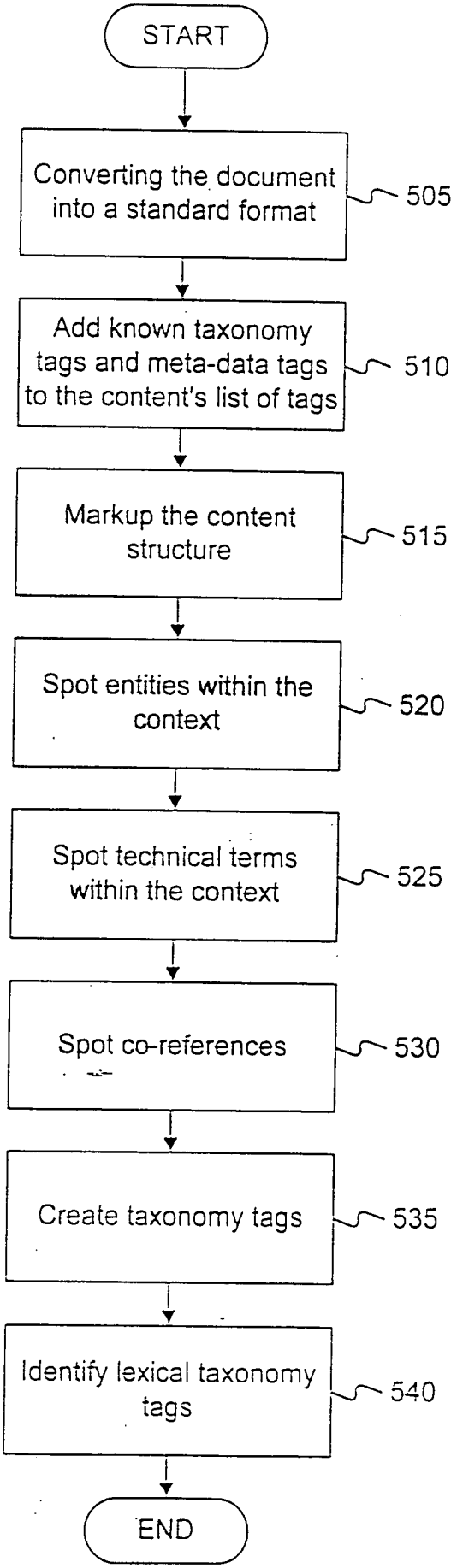

START
Converting the document into a standard format
505
Add known taxonomy tags and meta-data tags to the content's list of tags
510
Markup the content structure
515
Spot entities within the context
520
Spot technical terms within the context
525
Spot co-references
530
Create taxonomy tags
535
Identify lexical taxonomy tags
540
END

FIG. 5

Slice Representation

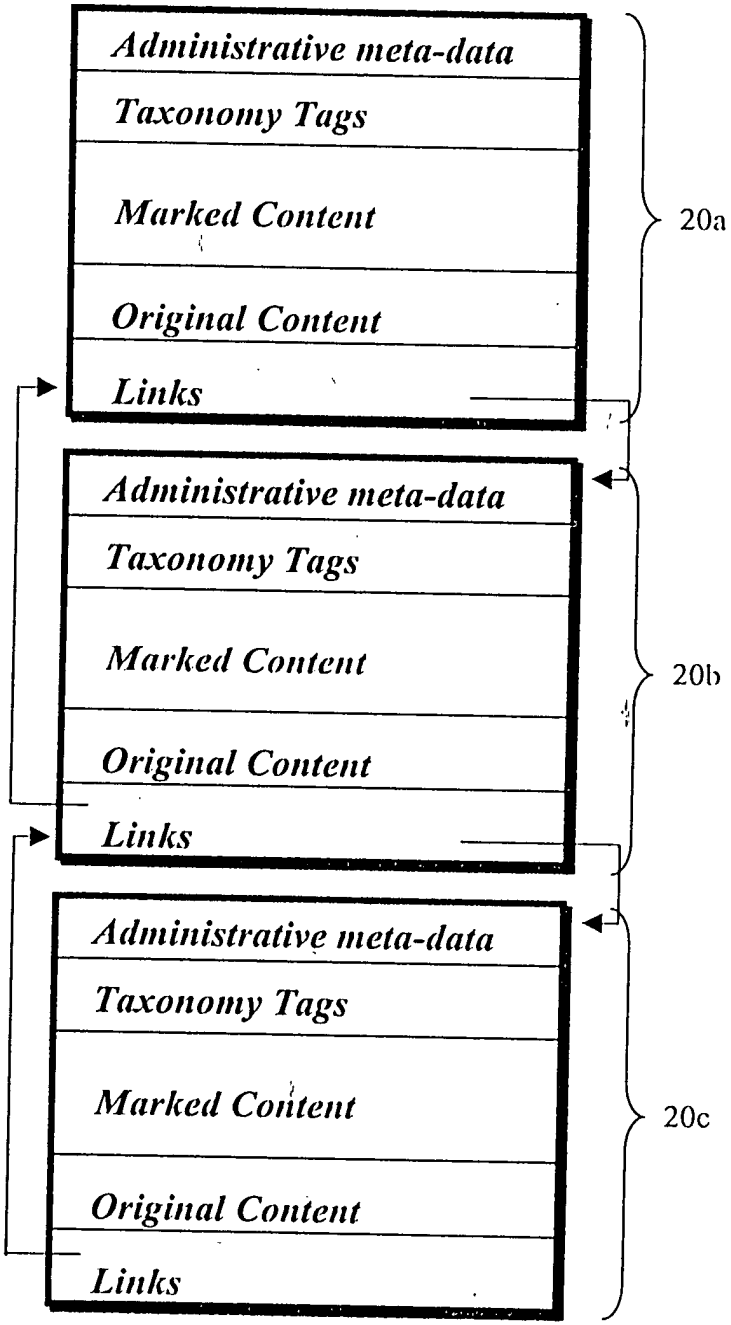
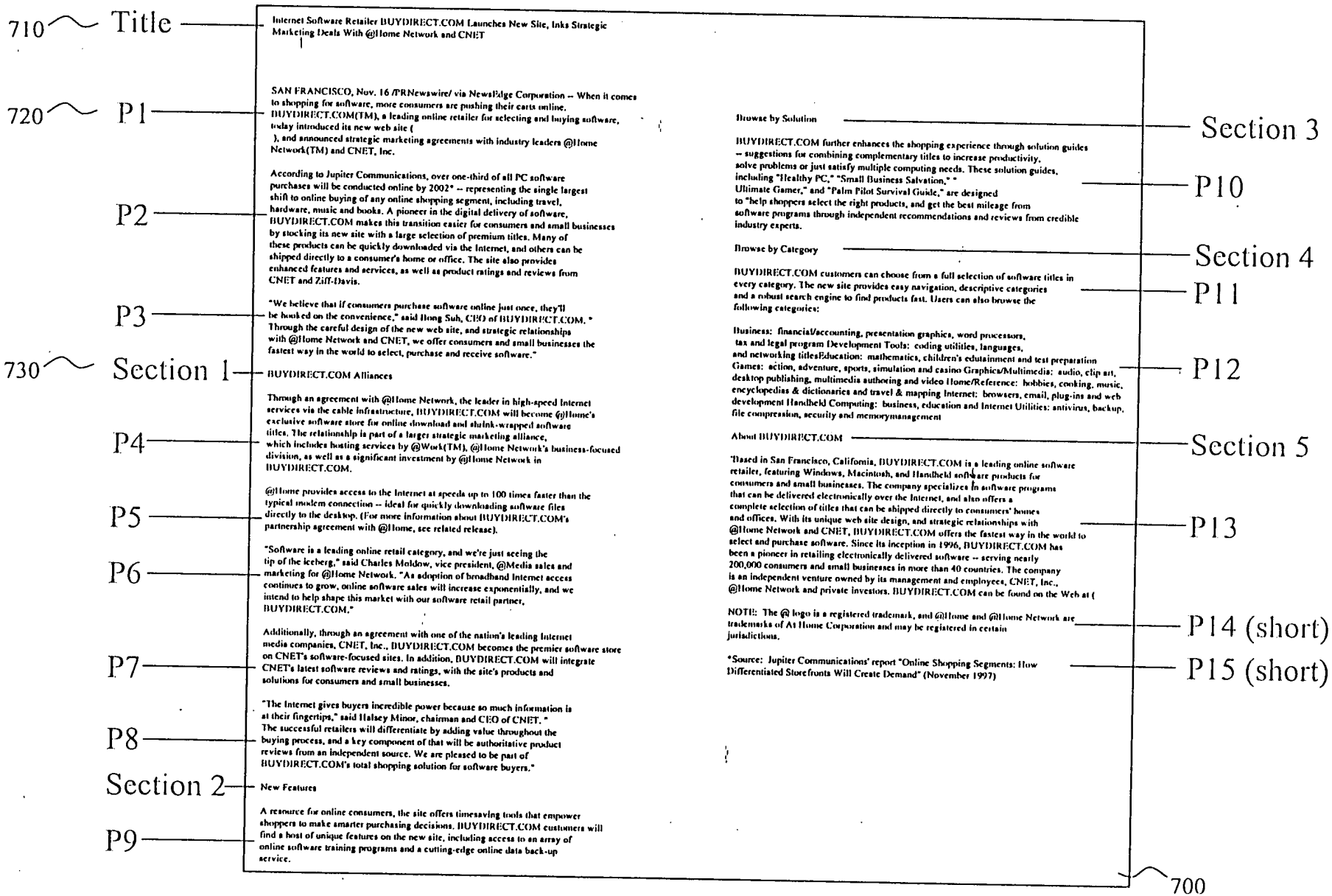


FIG. 6

Slicing (1)



Internet Software Retailer BUYDIRECT.COM Launches New Site, Inks Strategic Marketing Deals With @Home Network and CNET

SAN FRANCISCO, Nov. 16 /PRNewswire/ via NewsEdge Corporation -- When it comes to shopping for software, more consumers are pushing their carts online. BUYDIRECT.COM(TM), a leading online retailer for selecting and buying software, today introduced its new web site ( ), and announced strategic marketing agreements with industry leaders @Home Network(TM) and CNET, Inc.

According to Jupiter Communications, over one-third of all PC software purchases will be conducted online by 2002* -- representing the single largest shift to online buying of any online shopping segment, including travel, hardware, music and books. A pioneer in the digital delivery of software, BUYDIRECT.COM makes this transition easier for consumers and small businesses by stocking its new site with a large selection of premium titles. Many of these products can be quickly downloaded via the Internet, and others can be shipped directly to a consumer's home or office. The site also provides enhanced features and services, as well as product ratings and reviews from CNET and Ziff-Davis.

"We believe that if consumers purchase software online just once, they'll be hooked on the convenience," said Hong Suh, CEO of BUYDIRECT.COM. " Through the careful design of the new web site, and strategic relationships with @Home Network and CNET, we offer consumers and small businesses the fastest way in the world to select, purchase and receive software."

BUYDIRECT.COM Alliances

Through an agreement with @Home Network, the leader in high-speed Internet services via the cable infrastructure, BUYDIRECT.COM will become @Home's exclusive software store for online download and shrink-wrapped software titles. The relationship is part of a larger strategic marketing alliance, which includes hosting services by @Work(TM), @Home Network's business-focused division, as well as a significant investment by @Home Network in BUYDIRECT.COM.

@Home provides access to the Internet at speeds up to 100 times faster than the typical modem connection -- ideal for quickly downloading software files directly to the desktop. (For more information about BUYDIRECT.COM's partnership agreement with @Home, see related release).

"Software is a leading online retail category, and we're just seeing the tip of the iceberg," said Charles Moldow, vice president, @Media sales and marketing for @Home Network. "As adoption of broadband Internet access continues to grow, online software sales will increase exponentially, and we intend to help shape this market with our software retail partner, BUYDIRECT.COM."

Additionally, through an agreement with one of the nation's leading Internet media companies, CNET, Inc., BUYDIRECT.COM becomes the premier software store on CNET's software-focused sites. In addition, BUYDIRECT.COM will integrate CNET's latest software reviews and ratings, with the site's products and solutions for consumers and small businesses.

"The Internet gives buyers incredible power because so much information is at their fingertips," said Halsey Minor, chairman and CEO of CNET. " The successful retailers will differentiate by adding value throughout the buying process, and a key component of that will be authoritative product reviews from an independent source. We are pleased to be part of BUYDIRECT.COM's total shopping solution for software buyers."

New Features

A resource for online consumers, the site offers timesaving tools that empower shoppers to make smarter purchasing decisions. BUYDIRECT.COM customers will find a host of unique features on the new site, including access to an array of online software training programs and a cutting-edge online data back-up service.

Browse by Solution

BUYDIRECT.COM further enhances the shopping experience through solution guides -- suggestions for combining complementary titles to increase productivity, solve problems or just satisfy multiple computing needs. These solution guides, including "Healthy PC," "Small Business Salvation," " Ultimate Gamer," and "Palm Pilot Survival Guide," are designed to "help shoppers select the right products, and get the best mileage from software programs through independent recommendations and reviews from credible industry experts.

Browse by Category

BUYDIRECT.COM customers can choose from a full selection of software titles in every category. The new site provides easy navigation, descriptive categories and a robust search engine to find products fast. Users can also browse the following categories:

Business: financial/accounting, presentation graphics, word processors, tax and legal program Development Tools: coding utilities, languages, and networking titlesEducation: mathematics, children's edutainment and test preparation Games: action, adventure, sports, simulation and casino Graphics/Multimedia: audio, clip art, desktop publishing, multimedia authoring and video Home/Reference: hobbies, cooking, music, encyclopedias & dictionaries and travel & mapping Internet: browsers, email, plug-ins and web development Handheld Computing: business, education and Internet Utilities: antivirus, backup, file compression, security and memorymanagement

About BUYDIRECT.COM

"Based in San Francisco, California, BUYDIRECT.COM is a leading online software retailer, featuring Windows, Macintosh, and Handheld software products for consumers and small businesses. The company specializes in software programs that can be delivered electronically over the Internet, and also offers a complete selection of titles that can be shipped directly to consumers' homes and offices. With its unique web site design, and strategic relationships with @Home Network and CNET, BUYDIRECT.COM offers the fastest way in the world to select and purchase software. Since its inception in 1996, BUYDIRECT.COM has been a pioneer in retailing electronically delivered software -- serving nearly 200,000 consumers and small businesses in more than 40 countries. The company is an independent venture owned by its management and employees, CNET, Inc., @Home Network and private investors. BUYDIRECT.COM can be found on the Web at (

NOTE: The @ logo is a registered trademark, and @Home and @Home Network are trademarks of At Home Corporation and may be registered in certain jurisdictions.

*Source: Jupiter Communications' report "Online Shopping Segments: How Differentiated Storefronts Will Create Demand" (November 1997)

FIG. 7

Slicing (2)

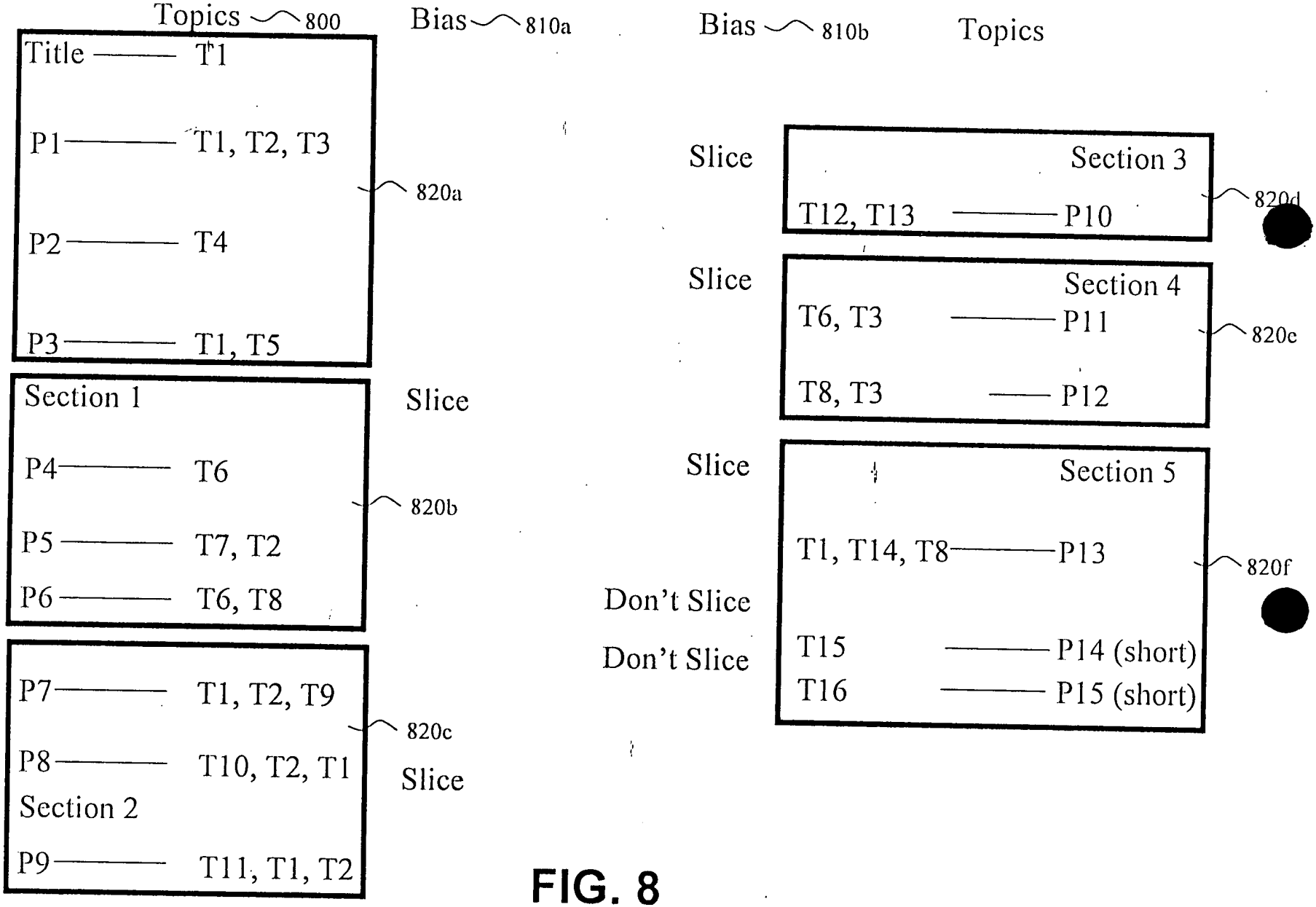


FIG. 8

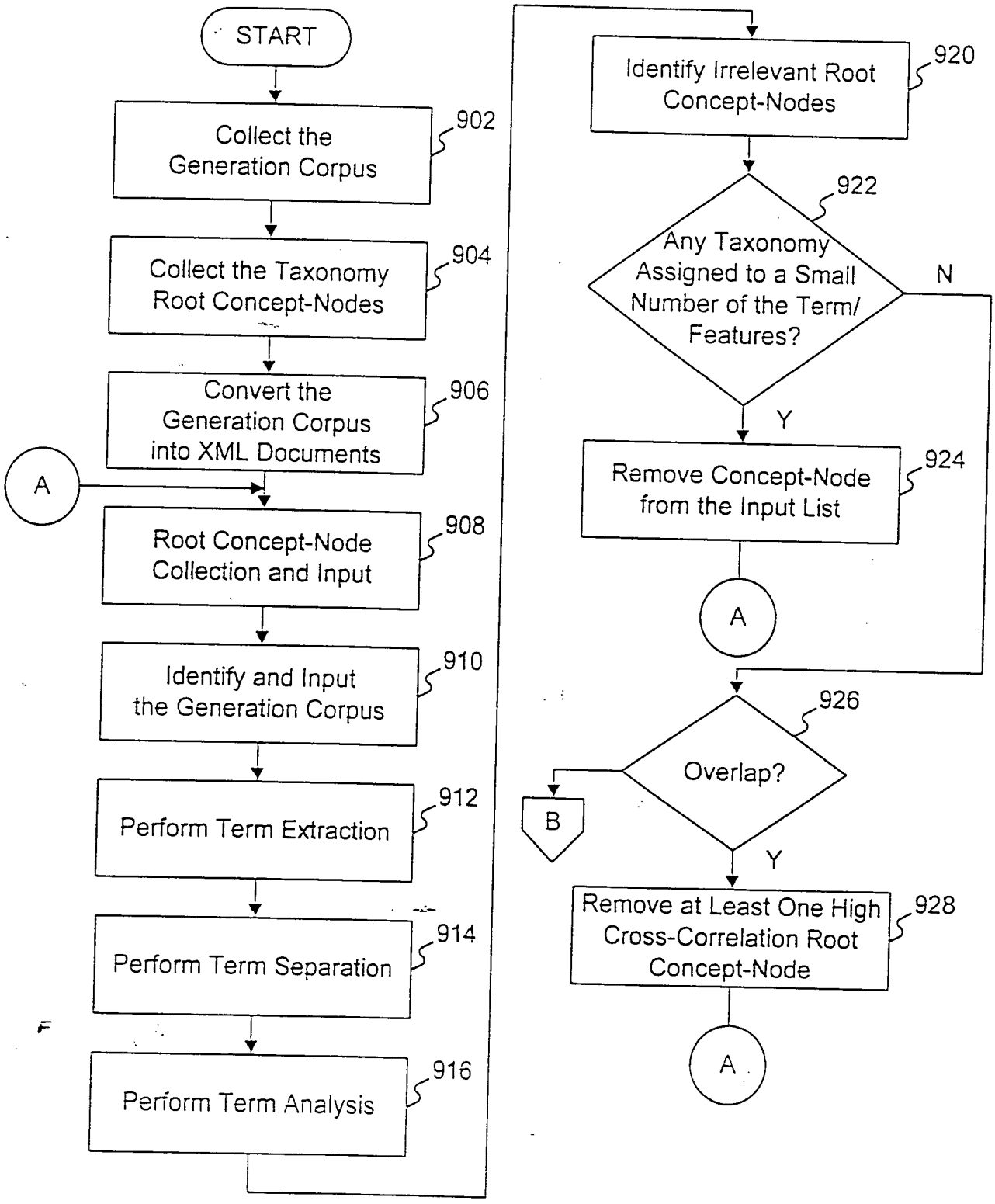

START
Collect the Generation Corpus
902
Collect the Taxonomy Root Concept-Nodes
904
Convert the Generation Corpus into XML Documents
906
A
Root Concept-Node Collection and Input
908
Identify and Input the Generation Corpus
910
Perform Term Extraction
912
Perform Term Separation
914
Perform Term Analysis
916
Identify Irrelevant Root Concept-Nodes
920
Any Taxonomy Assigned to a Small Number of the Term/ Features?
922
N
Y
Remove Concept-Node from the Input List
924
A
Overlap?
926
B
Y
Remove at Least One High Cross-Correlation Root Concept-Node
928
A

FIG. 9a

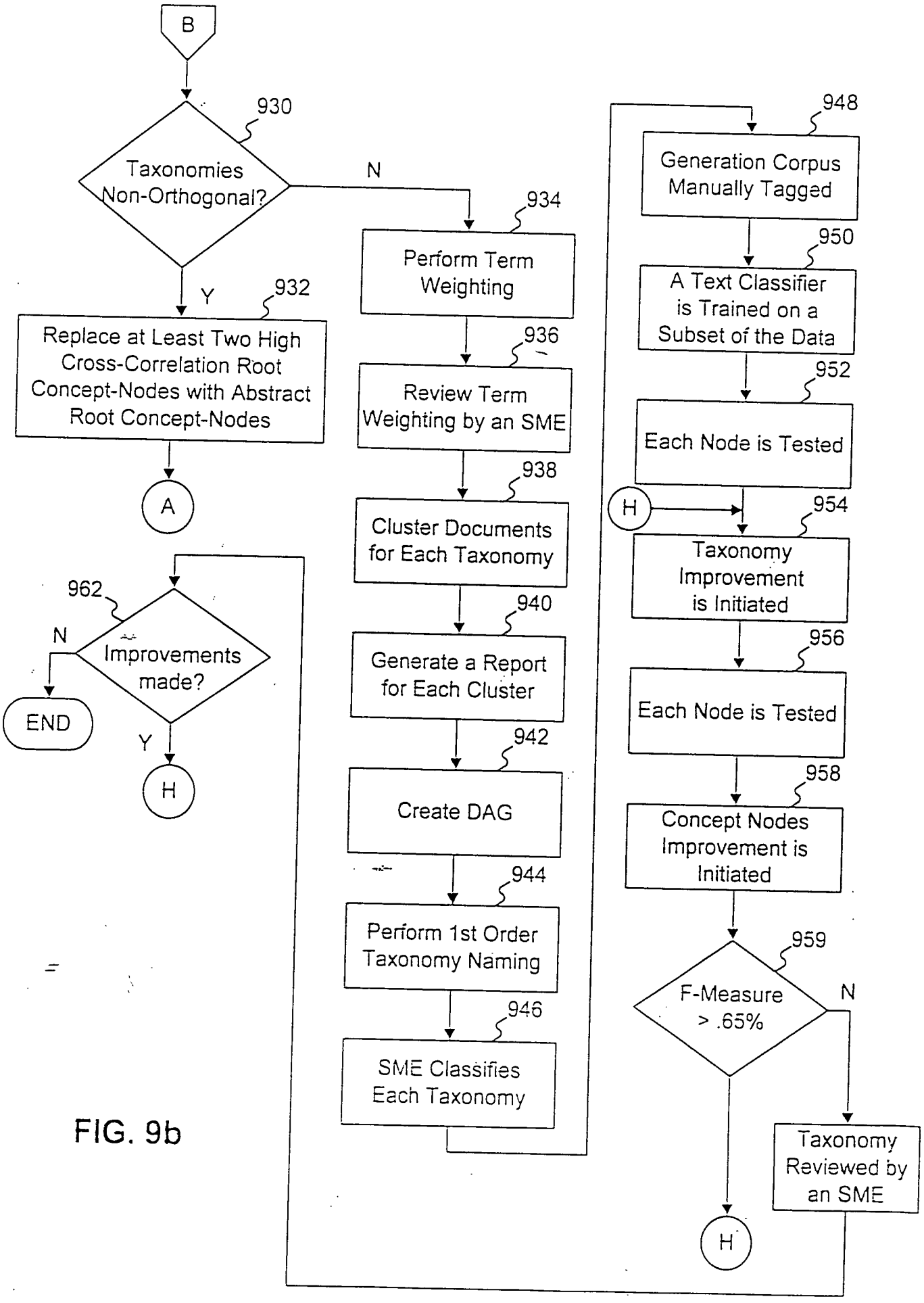

B
930
Taxonomies Non-Orthogonal?
N
Y
932
Replace at Least Two High Cross-Correlation Root Concept-Nodes with Abstract Root Concept-Nodes
A
962
N
Improvements made?
END
Y
H
934
Perform Term Weighting
936
Review Term Weighting by an SME
938
Cluster Documents for Each Taxonomy
940
Generate a Report for Each Cluster
942
Create DAG
944
Perform 1st Order Taxonomy Naming
946
SME Classifies Each Taxonomy
948
Generation Corpus Manually Tagged
950
A Text Classifier is Trained on a Subset of the Data
952
Each Node is Tested
H
954
Taxonomy Improvement is Initiated
956
Each Node is Tested
958
Concept Nodes Improvement is Initiated
959
F-Measure > .65%
N
Taxonomy Reviewed by an SME
H

FIG. 9b

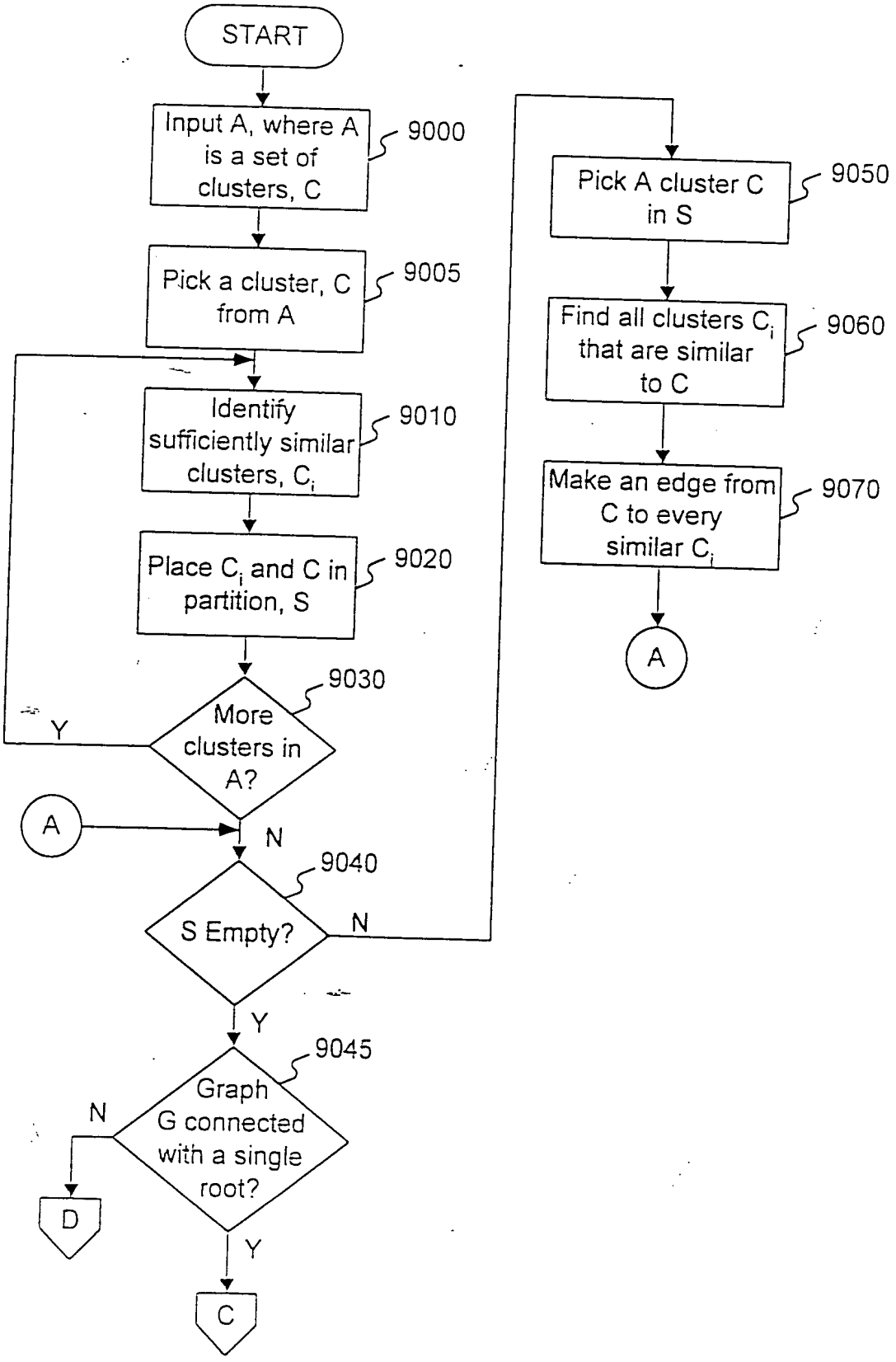

START
Input A, where A is a set of clusters, C
9000
Pick a cluster, C from A
9005
Identify sufficiently similar clusters, C_i
9010
Place C_i and C in partition, S
9020
More clusters in A?
9030
Y
N
A
S Empty?
9040
N
Y
Graph G connected with a single root?
9045
N
Y
D
C
Pick A cluster C in S
9050
Find all clusters C_i that are similar to C
9060
Make an edge from C to every similar C_i
9070
A

FIG. 9c

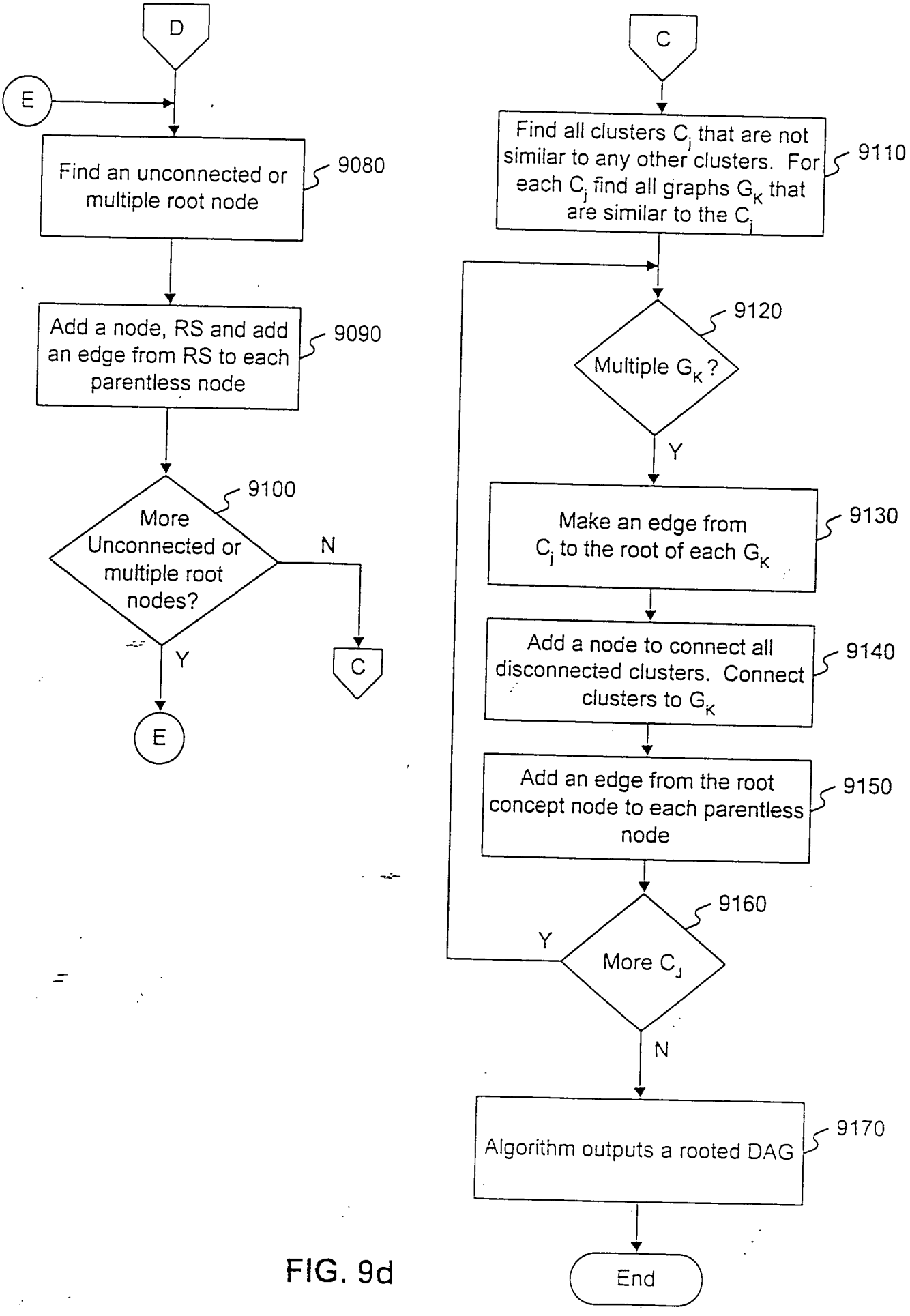

D
E
Find an unconnected or multiple root node
9080
Add a node, RS and add an edge from RS to each parentless node
9090
More Unconnected or multiple root nodes?
9100
N
Y
C
E
C
Find all clusters C_j that are not similar to any other clusters. For each C_j find all graphs G_K that are similar to the C_j
9110
Multiple G_K ?
9120
Y
Make an edge from C_j to the root of each G_K
9130
Add a node to connect all disconnected clusters. Connect clusters to G_K
9140
Add an edge from the root concept node to each parentless node
9150
More C_J
9160
Y
N
Algorithm outputs a rooted DAG
9170
End

FIG. 9d

A Taxonomy of Document Sources

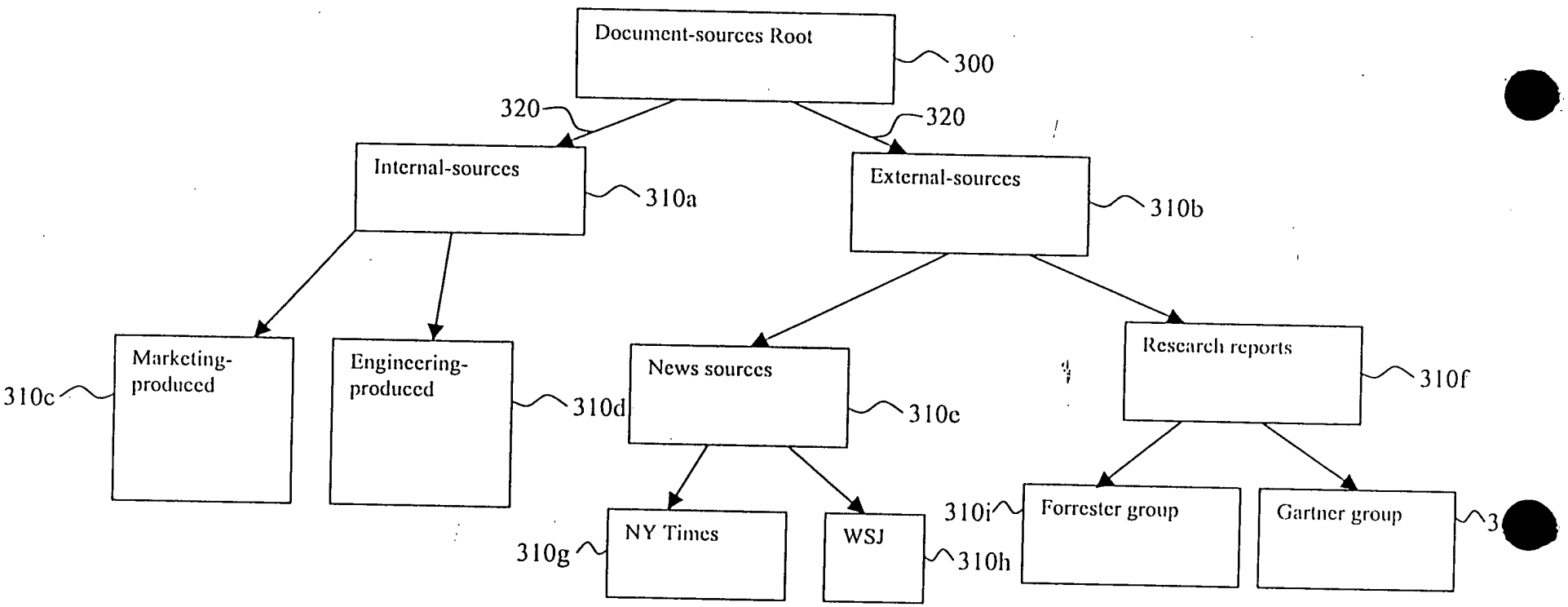


FIG. 10

30a

Audience Taxonomy

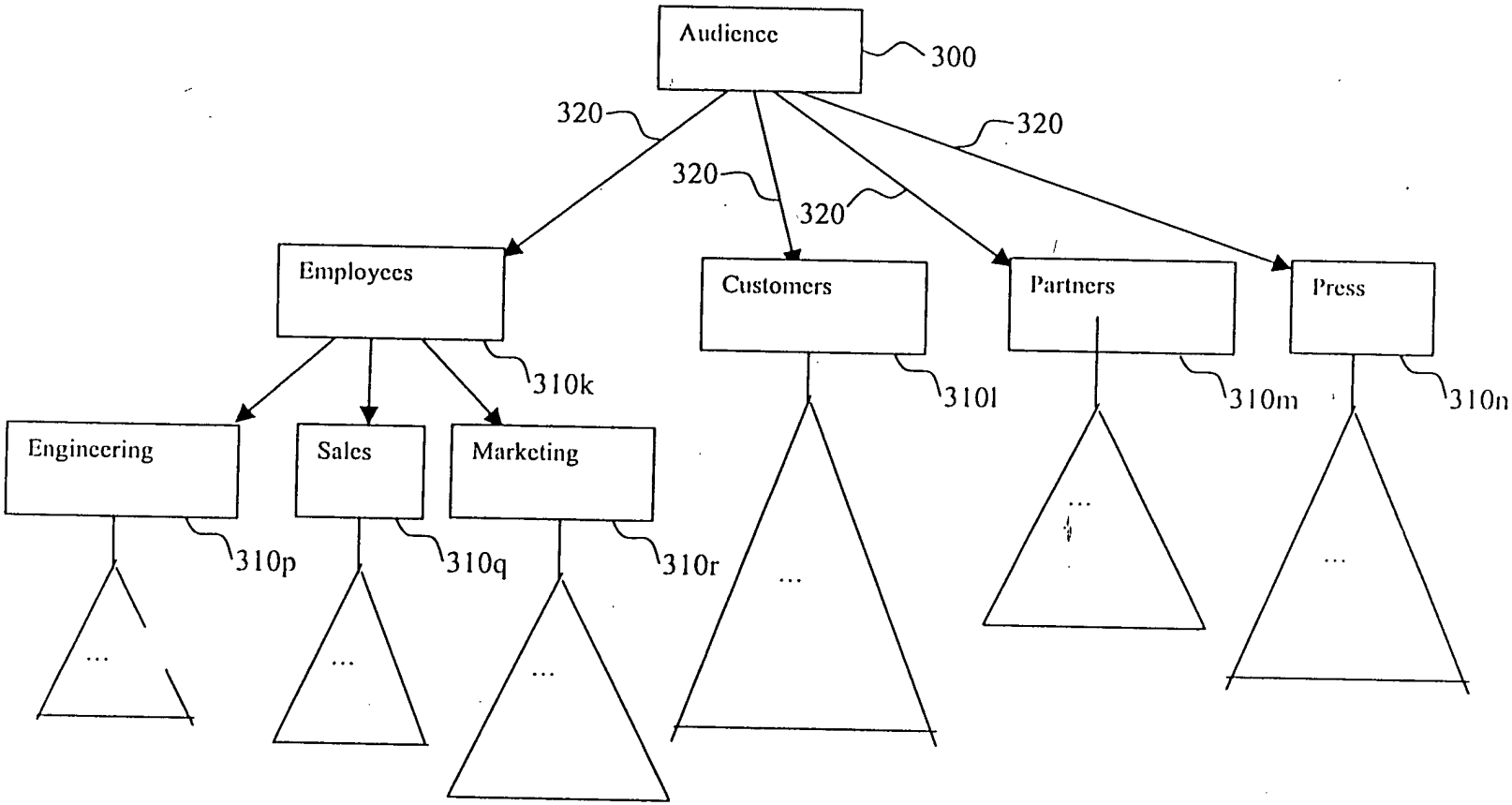
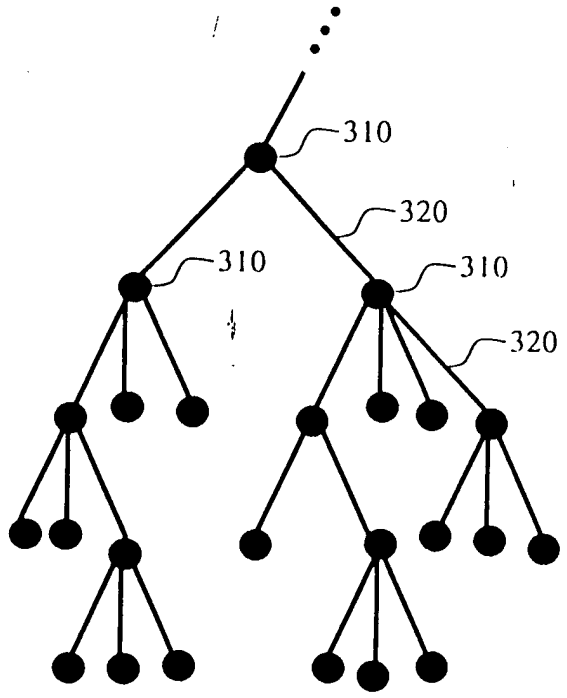


FIG. 11

Index Preparation---Step 1

For each node:

Make a "node document" which is the concatenation of all documents tagged to the node (here represented by ●).



30

FIG. 12

Index Preparation---Step 2

Cluster the nodes according to similarity in vocabulary usage in the node-documents. Each lined-pattern represents a cluster.

- orange
- purple
- green
- blue

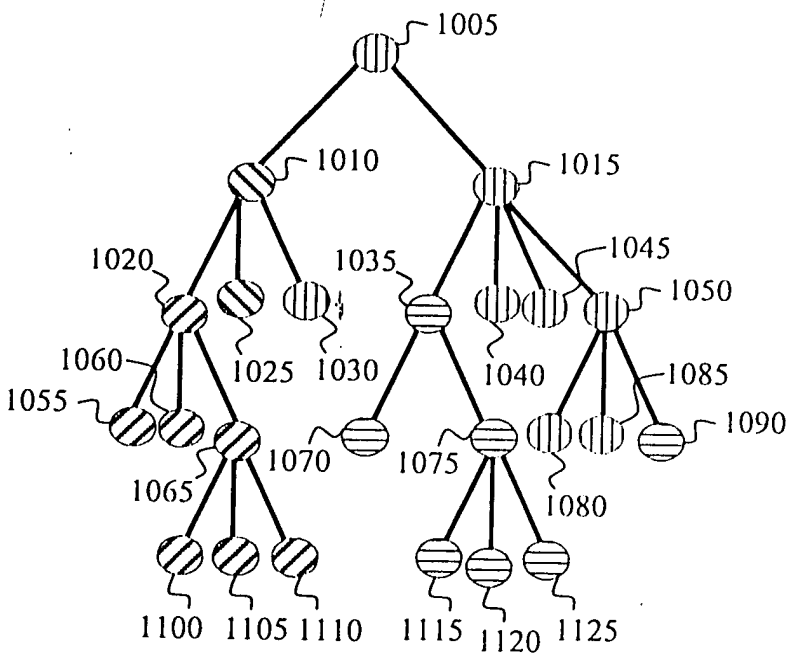


FIG. 13

Index Preparation---Step 3

For each cluster,
Build an index over the documents tagged to nodes in the cluster.

- orange
- purple
- green
- blue

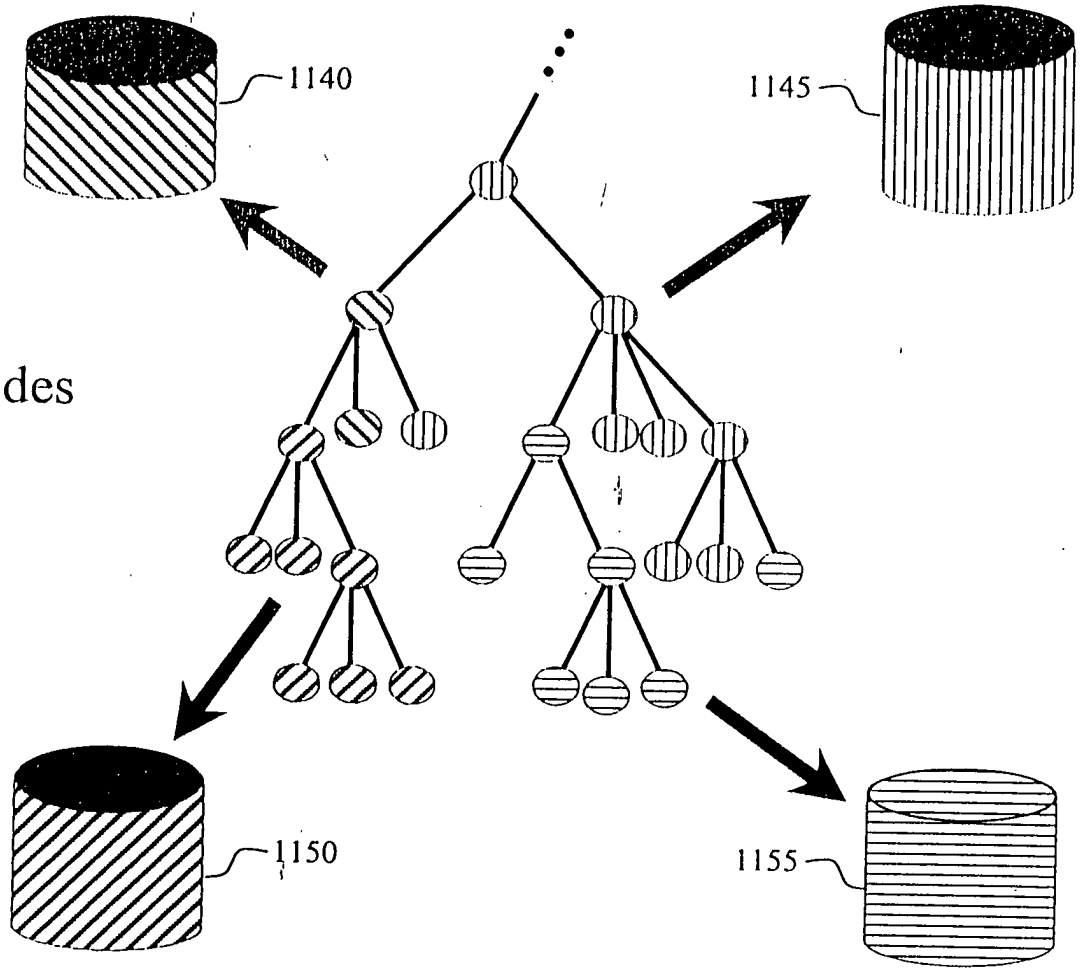


FIG. 14

Region Designation: Marking

Identify nodes within distance *D* of a number *N* of query taxonomy tags. A node so identified is a *marked* node. Illustratively, here *D* is one and *N* is two.

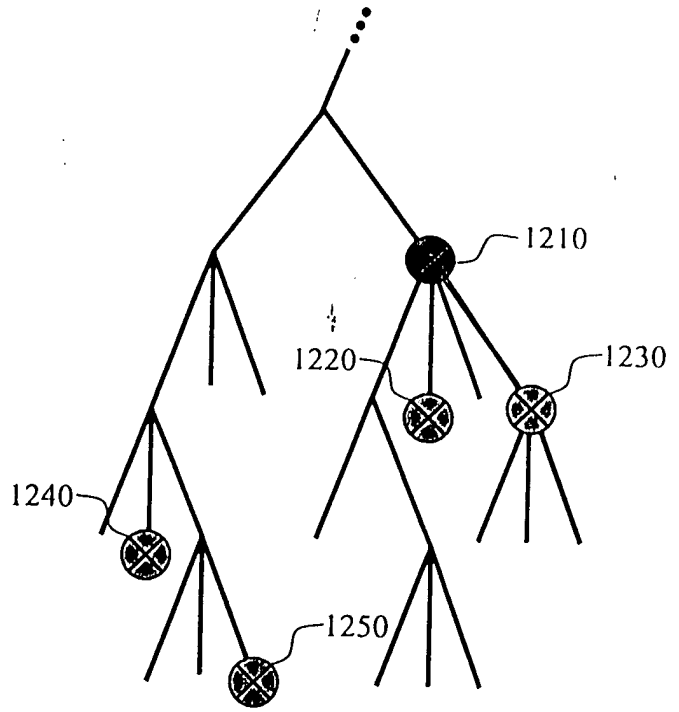


FIG. 15

Region Designation: Smoothing

Optionally, identify nodes within distance 1 of a query taxonomy tag or marked node. A node so identified is a *smoothed* node.

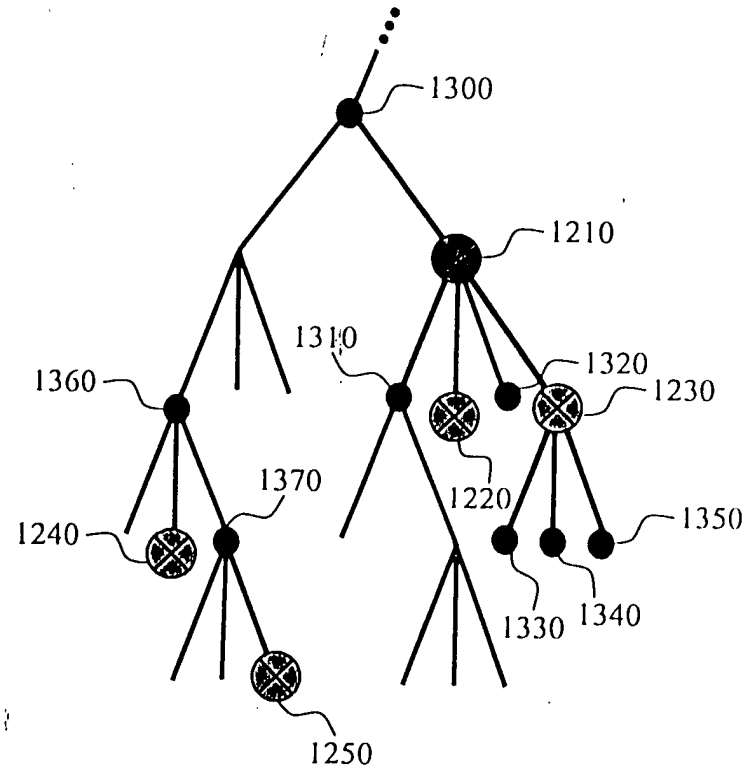


FIG. 16

Region Designation: Aggregation

Identify groups of query taxonomy tags, marked nodes, and smoothed nodes that, transitively, are within distance *D'* from each other. Illustratively, here *D'* is one.

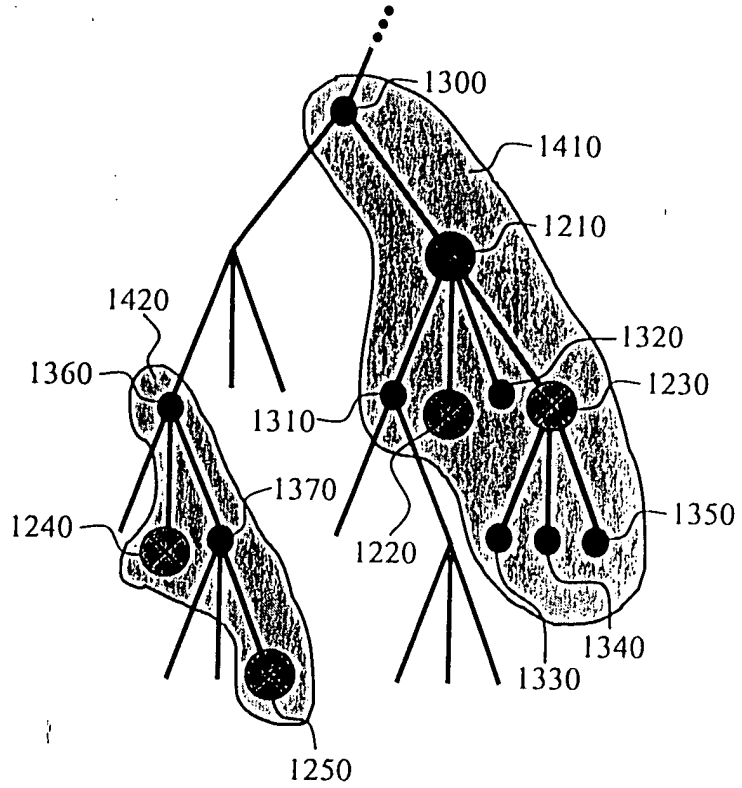


FIG. 17

Search: Index Identification

For each region,
Identify a set of indexes that covers all of the nodes in the region.

- orange
- purple
- green
- blue

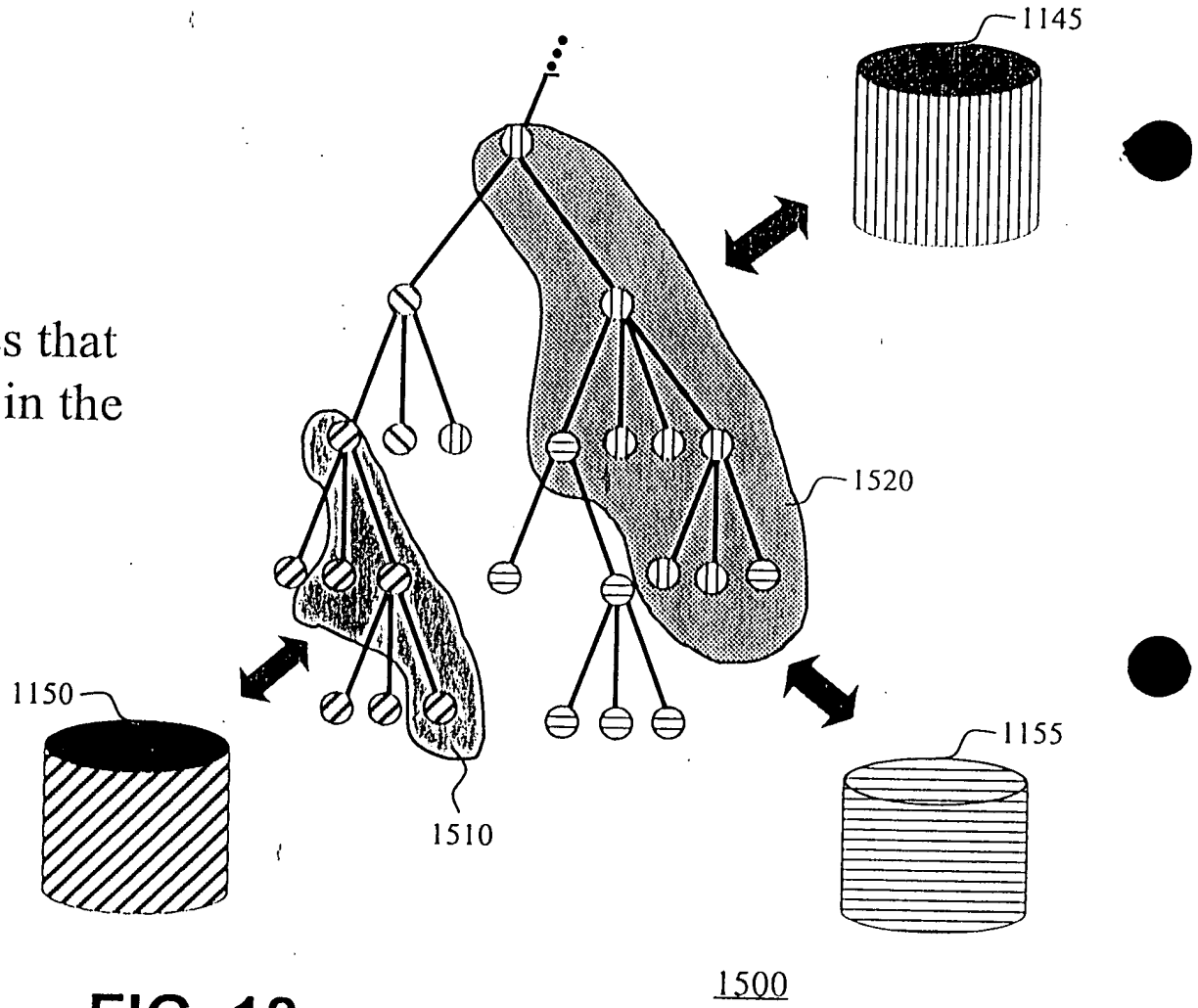


FIG. 18

Ranking: Search Engine Rank

For each document the rank returned by the search engine is adjusted by ...

For purposes of illustration, rank is shown by distance from the bottom of the picture.

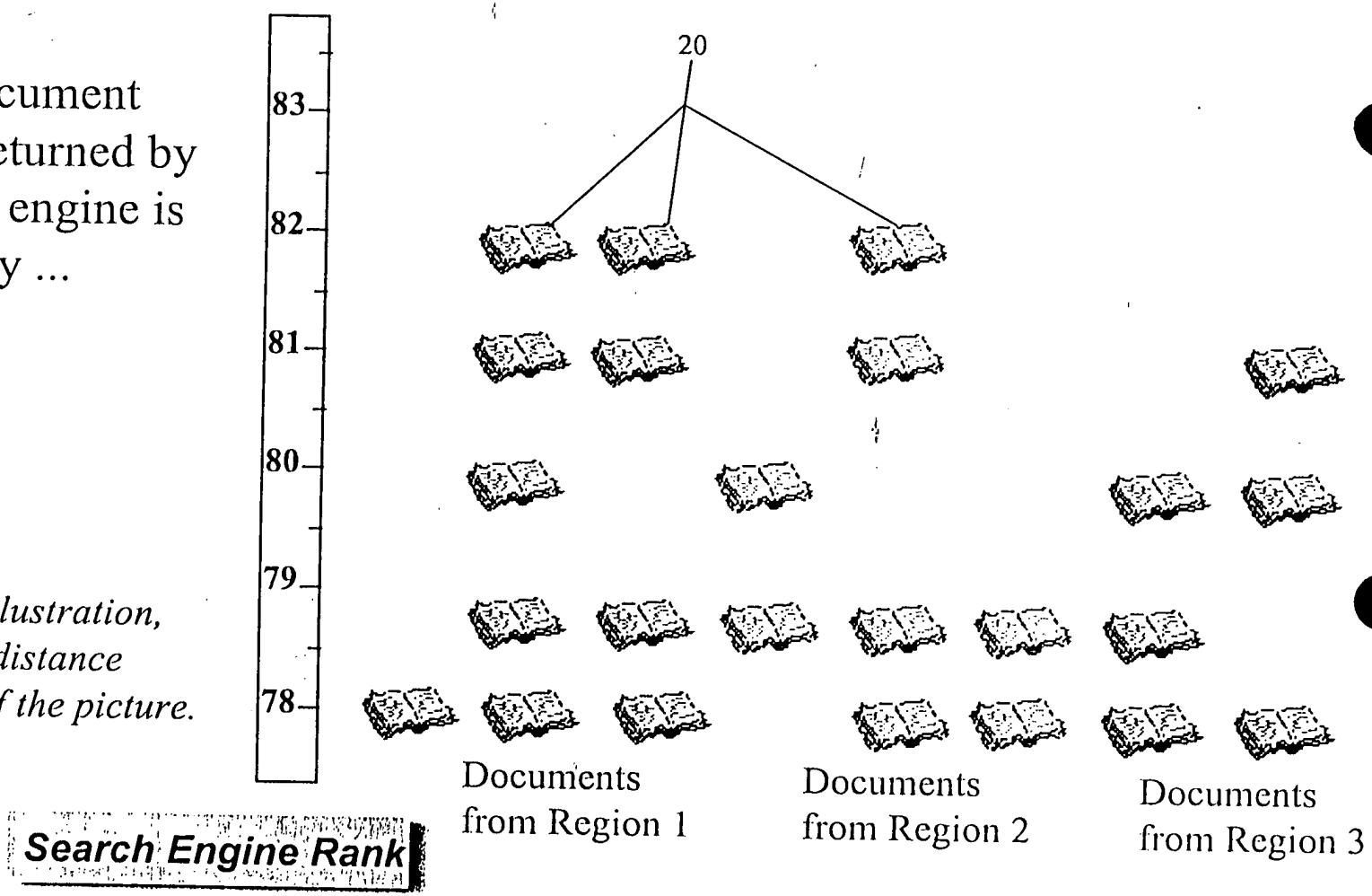
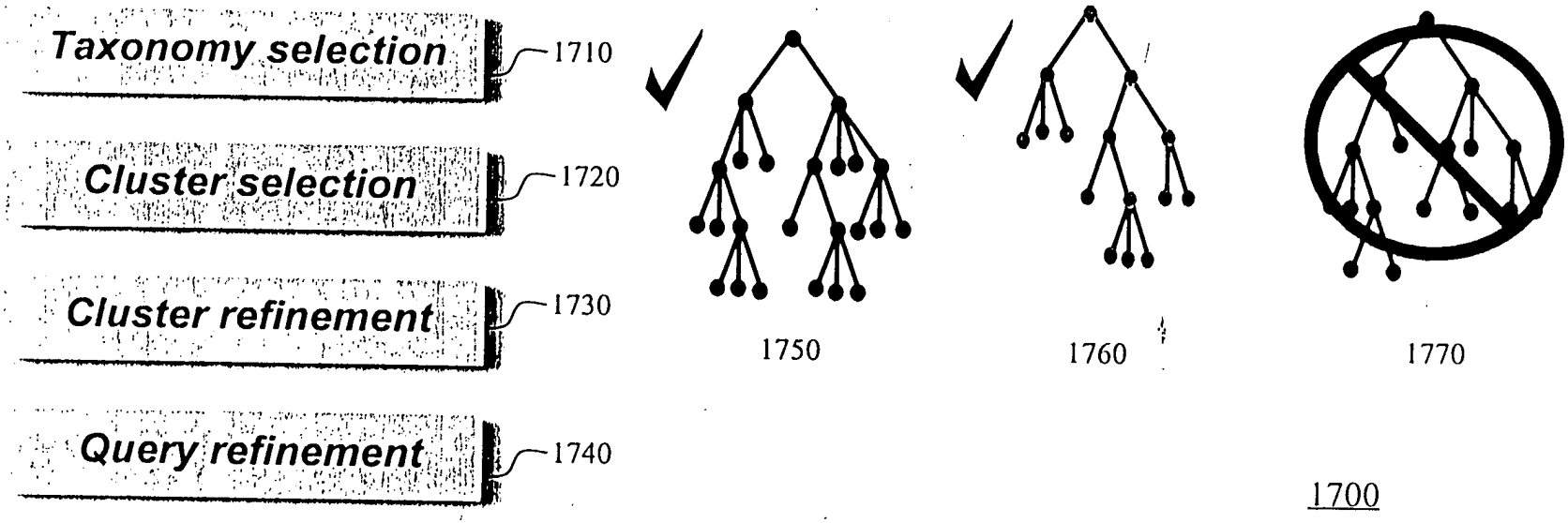


FIG. 19

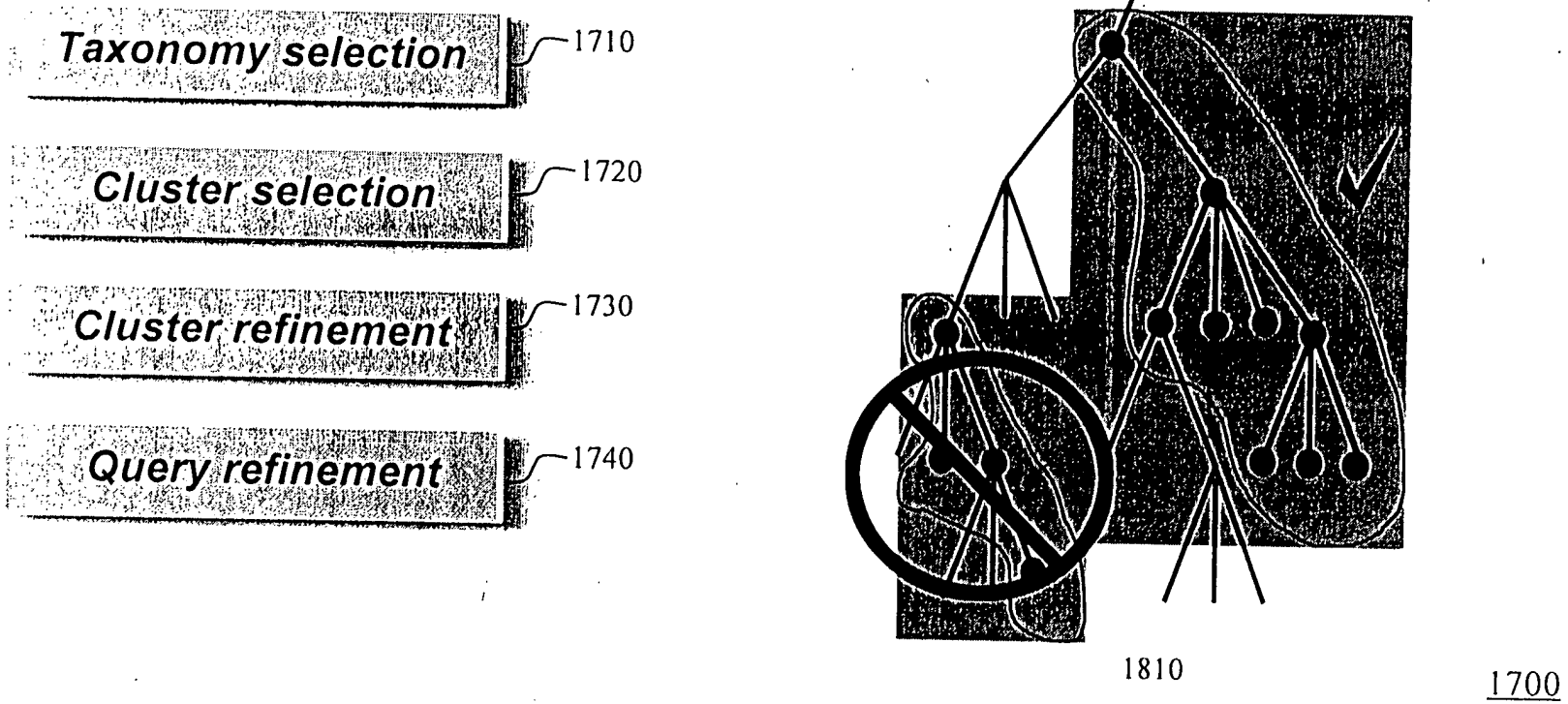
Interactive Dialogue



The user can choose among the taxonomies.

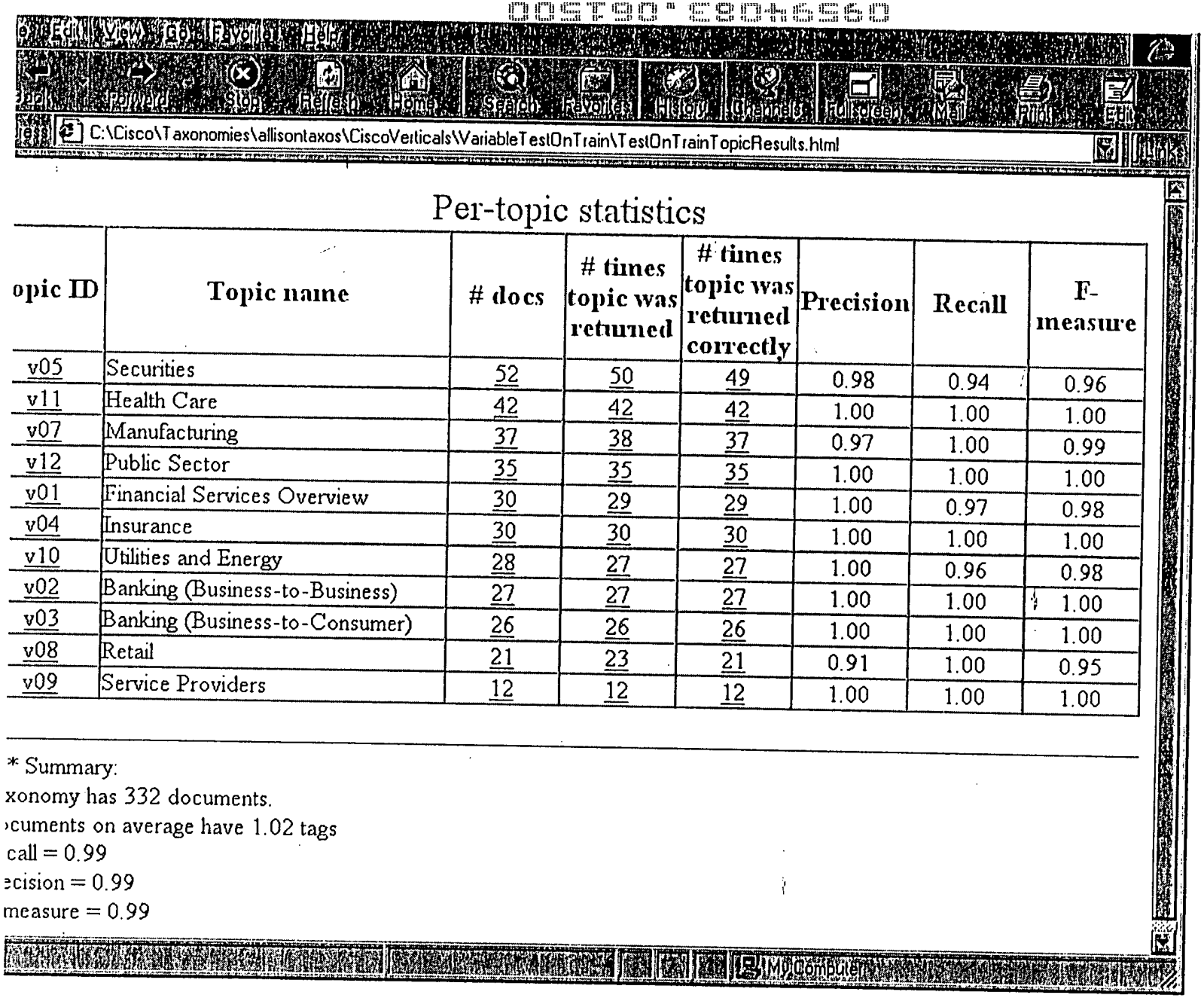
FIG. 20

Interactive Dialogue



The user can choose among the clusters.

FIG. 21



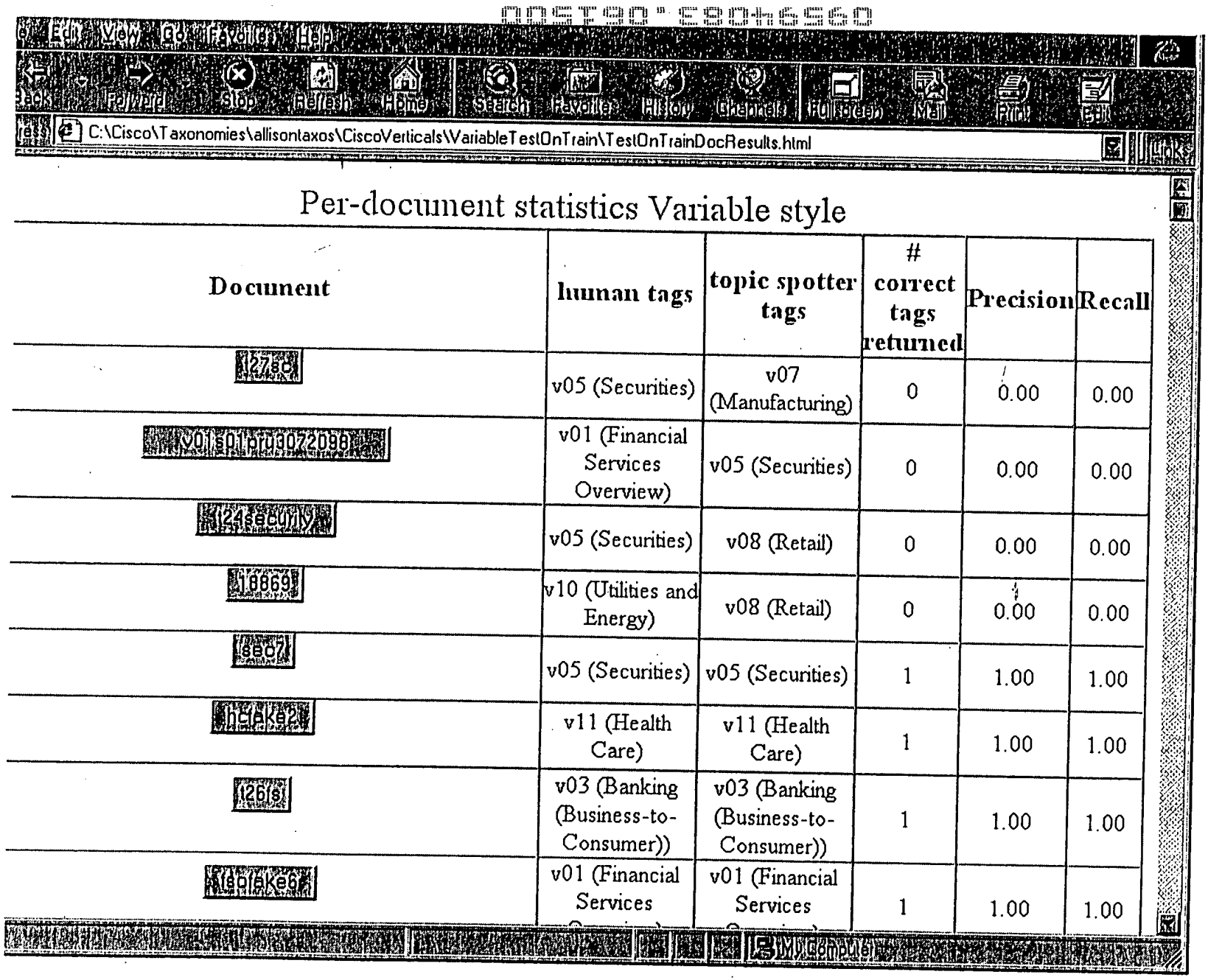
Per-topic statistics

opic ID	Topic name	# docs	# times topic was returned	# times topic was returned correctly	Precision	Recall	F-measure
v05	Securities	52	50	49	0.98	0.94	0.96
v11	Health Care	42	42	42	1.00	1.00	1.00
v07	Manufacturing	37	38	37	0.97	1.00	0.99
v12	Public Sector	35	35	35	1.00	1.00	1.00
v01	Financial Services Overview	30	29	29	1.00	0.97	0.98
v04	Insurance	30	30	30	1.00	1.00	1.00
v10	Utilities and Energy	28	27	27	1.00	0.96	0.98
v02	Banking (Business-to-Business)	27	27	27	1.00	1.00	1.00
v03	Banking (Business-to-Consumer)	26	26	26	1.00	1.00	1.00
v08	Retail	21	23	21	0.91	1.00	0.95
v09	Service Providers	12	12	12	1.00	1.00	1.00

* Summary:
xonomy has 332 documents.
›cuments on average have 1.02 tags
call = 0.99
ecision = 0.99
measure = 0.99

Test On Train Report 1

FIG. 22

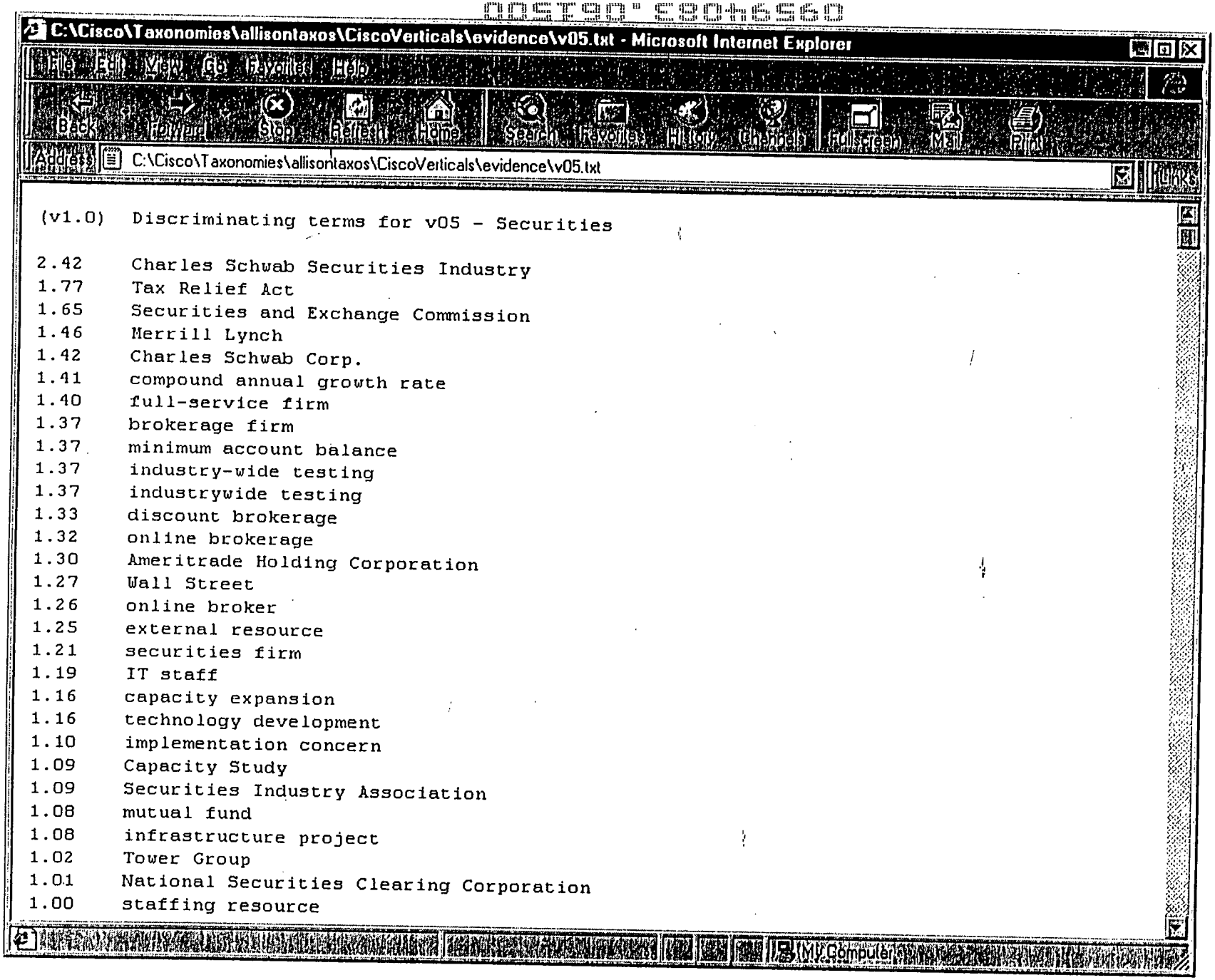


Per-document statistics Variable style

Document	human tags	topic spotter tags	# correct tags returned	Precision	Recall
12736	v05 (Securities)	v07 (Manufacturing)	0	0.00	0.00
[illegible]	v01 (Financial Services Overview)	v05 (Securities)	0	0.00	0.00
124secuty	v05 (Securities)	v08 (Retail)	0	0.00	0.00
18869	v10 (Utilities and Energy)	v08 (Retail)	0	0.00	0.00
seo7	v05 (Securities)	v05 (Securities)	1	1.00	1.00
[illegible]	v11 (Health Care)	v11 (Health Care)	1	1.00	1.00
126fs	v03 (Banking (Business-to-Consumer))	v03 (Banking (Business-to-Consumer))	1	1.00	1.00
[illegible]	v01 (Financial Services	v01 (Financial Services	1	1.00	1.00

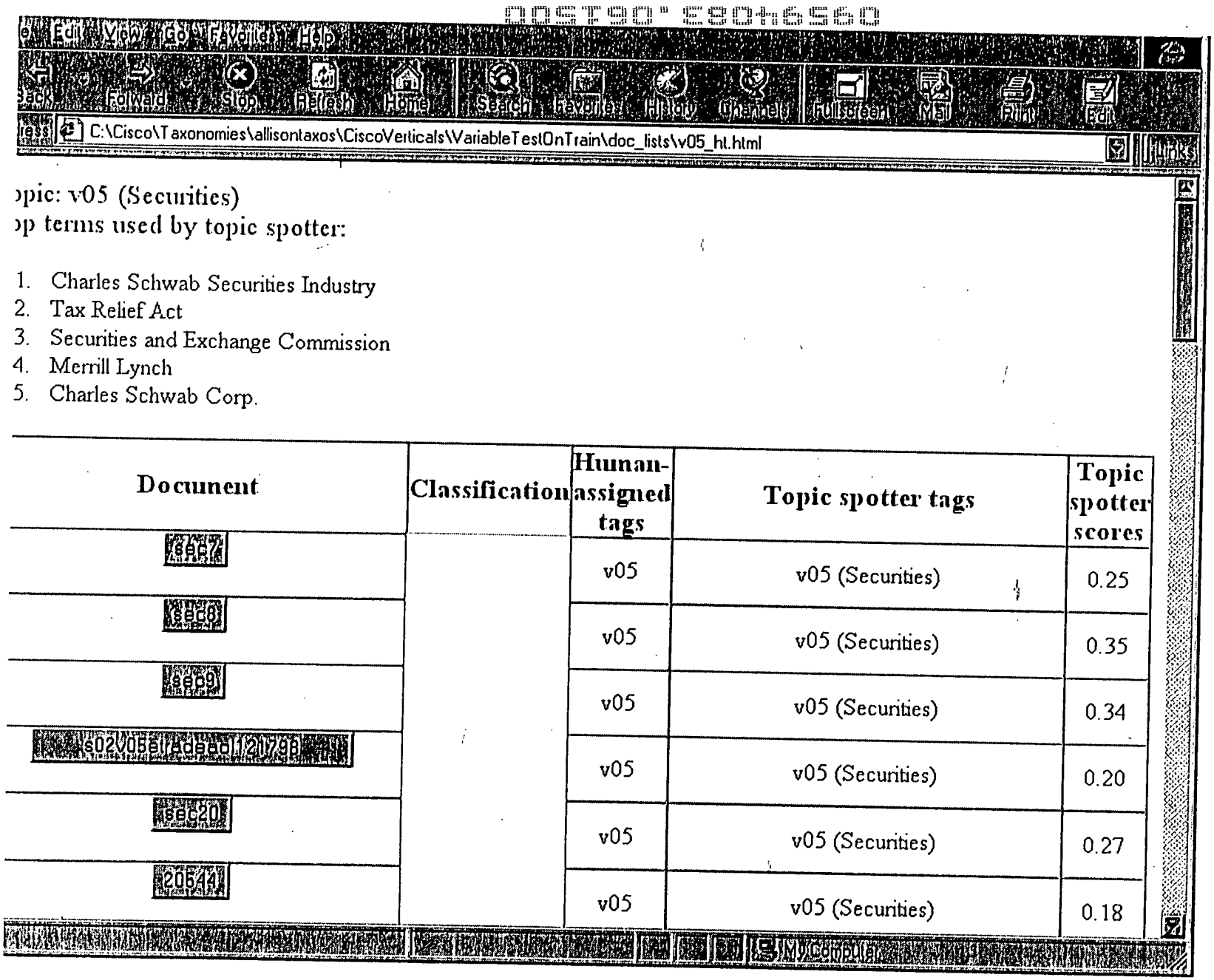
Test On Train Report 2

FIG. 23



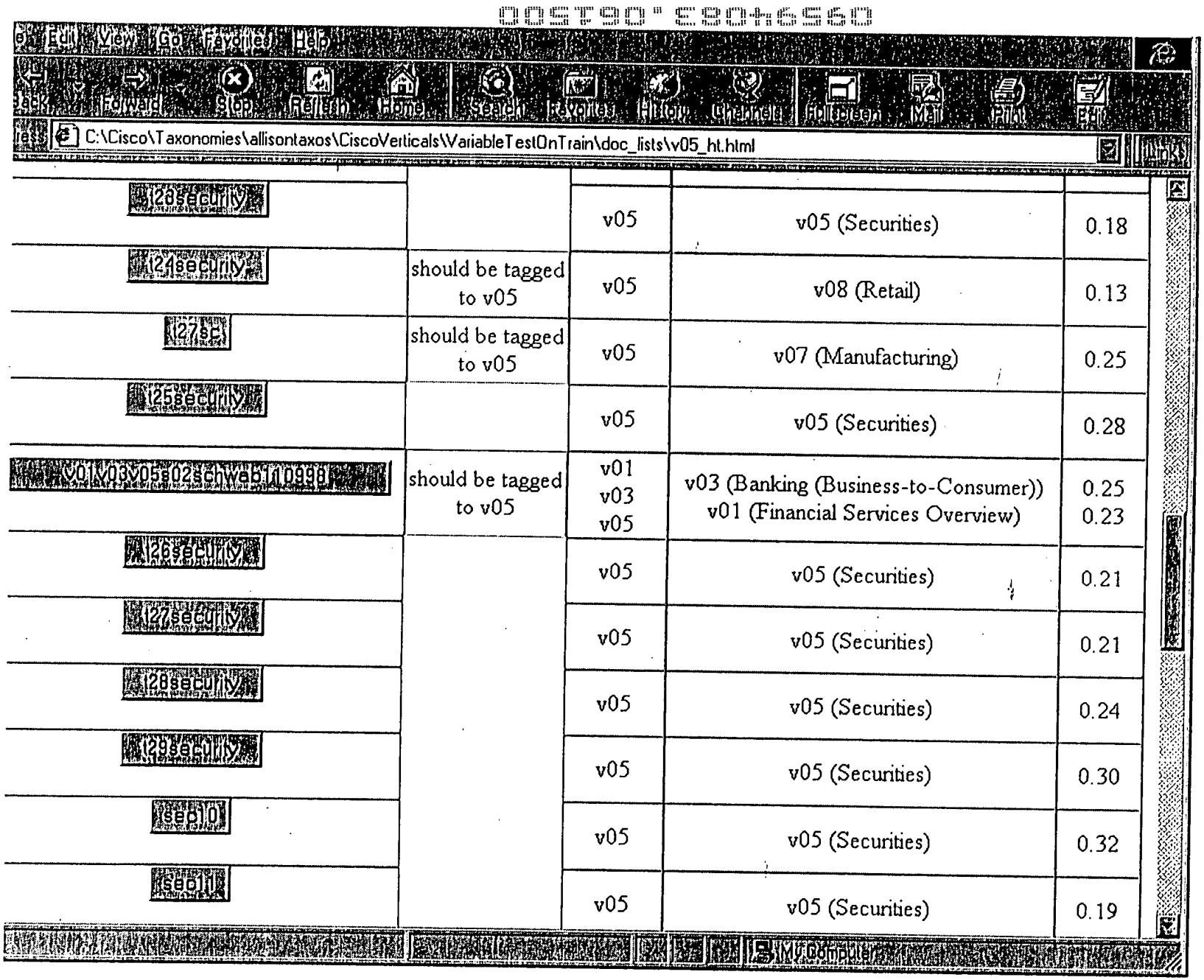
Test On Train Report 3

FIG. 24



Test On Train Report 4

FIG. 25



i28security		v05	v05 (Securities)	0.18
i24security	should be tagged to v05	v05	v08 (Retail)	0.13
i27sc	should be tagged to v05	v05	v07 (Manufacturing)	0.25
i25security		v05	v05 (Securities)	0.28
v01v03v05s02schwab1n998	should be tagged to v05	v01 v03 v05	v03 (Banking (Business-to-Consumer)) v01 (Financial Services Overview)	0.25 0.23
i26security		v05	v05 (Securities)	0.21
i27security		v05	v05 (Securities)	0.21
i28security		v05	v05 (Securities)	0.24
i29security		v05	v05 (Securities)	0.30
sec10		v05	v05 (Securities)	0.32
sec11		v05	v05 (Securities)	0.19

Test On Train Report 5

FIG. 26